Learning Thematic Role Relations for Wordnets

Andreas Wagner

Abstract

In this paper, I present a method for learning thematic role relations (selectional preferences) for wordnets by means of statistical corpus analysis. An evaluation on a gold standard, which I extracted from EuroWordNet, shows that this method achieves a learning accuracy of up to 77%. I also propose a preprocessing step for a partial lexical disambiguation of the input data. This disambiguation step reduces noise in the output of the learning algorithm, accompanied by a comparably moderate loss in performance.

1 Introduction

Broad-coverage, general-purpose lexical semantic knowledge bases like WordNet (cf. [Fellbaum, 1998]) are avaliable and/or developed for a variety of languages. For example, EuroWordNet (cf. [Vossen, 1999]) is a collection of wordnets for eight European languages, which are aligned via an inter-lingual index. These wordnets consist of concept nodes which represent word senses as synonym sets (synsets) and semantic relations between these concept nodes, e.g. hyperonymy, antonymy, or holonymy. Such wordnets are widely used as background knowledge for a number of HLT tasks. The concepts provide the inventory for word sense disambiguation (WSD) and semantic annotation of corpora. The relations between concepts provide useful information e.g. for measuring semantic similarity for WSD or text analysis, query expansion in information retrieval, or semantic inferencing.

One group of semantic relations in EuroWordNet are thematic role relations. These relations connect verbal concepts with nominal concepts which preferably occur as their complements. For example, the verbal concept <eat> should have AGENT pointers to the nominal concepts <human> and <animal>, and a PATIENT pointer to the concept <food>. Thematic relations provide information about the selectional preferences which verbs impose on their complements. In addition to the tasks mentioned above, this kind of information is useful for syntactic disambiguation (cf. [Resnik, 1993], [Abe and Li, 1996]).

As manual encoding of lexical knowledge is particularly labour-intensive, (semi-)automatic methods for learning lexical information have been explored. This paper addresses the acquisition of selectional preferences by means of statistical corpus analysis. In particular, the approach described here aims at learning selectional preferences in a form that is suitable for integration in wordnets, namely as thematic role relations. This implies that the acquired selectional preferences have to

(a) be represented as relations between verb concepts and noun concepts (as opposed to relations between verb forms and noun forms or noun concepts)

(b) represent the appropriate level of generalization.

The latter point means that the noun concepts which are acquired as preferences of a certain verb concept are at a level of abstraction that is both *empirically adequate* (i.e. captures all and only the preferred concepts) and *as compact as possible*. For example, it is inappropriate to introduce PATIENT relations from <eat> to a concept which is so general that it also subsumes dispreferred concepts (e.g. <entity>). On the other hand, it is not desirable to establish relations to all the food concepts in the wordnet (<potato>, <milk chocolate>, <mock turtle soup>,...) because this would be highly redundant and would not express any generalization. One would rather want to find a concept which just subsumes all the preferred concepts (such as <food>). The desideratum to acquire concepts which are not too specific is motivated from a practical point of view (storage economy) as well as by conceptual considerations (appropriate generalizations should be expressed; this is important for applications like semantic inferencing).

It is important to point out that these general considerations should be regarded relative to the needs of a particular application and/or approach. For example, there are parsing systems (e.g. [Huyck, 2000]) which make use of selectional restrictions modeled by a small set of very abstract semantic classes (like *human*) to resolve structural ambiguities. On the other hand, there are lexical resources (e.g. FrameNet, c.f. [Fillmore et al., 2001]) where selectional information is represented by nouns which *typically* occur at a certain frame slot of a particular verb. Thus, these resources model selectional preferences at a very low level of generalization. This may be appropriate for particular HLT applications, e.g. in a text understanding system tuned for a particular domain (c.f. [Fillmore and Baker, 2001]). A reusable approach for acquiring selectional preferences should be adjustable to such *a priori* design principles with respect to the desired degree of generalization, and, within this degree, behave as sketched in the previous paragraph.

2 Acquiring selectional preferences from corpora

Several approaches for acquiring selectional preferences by statistical corpus analysis based on WordNet have been proposed (e.g., cf. [Resnik, 1993], [Ribas, 1994], [Li and Abe, 1995], [Abe and Li, 1996], [Agirre and Martinez, 2002]). However, to our knowledge, none of these approaches fulfil both criteria (a) and (b) mentioned in the last section.

Concerning (a), most of the methods mentioned above learn relations between verb *forms* and noun concepts. Only [Agirre and Martinez, 2002] acquire preferences between verb and noun concepts. Concerning (b), [Resnik, 1993] does not aim at characterizing selectional preferences of a verb by a "representative" subset of concepts, but keeps all WordNet noun concepts (and the corresponding preference values) for each verb. [Ribas, 1994] and [Agirre and Martinez, 2002] identify a set of "representative concepts" by a simple greedy algorithm. The impact of this algorithm for the generalization level of the selected concepts is undetermined. Only the work of Li & Abe systematically addresses the problem of appropriate generalization. Li & Abe obtain a set of concepts that form a partition of the corpus instances. They employ a theoretically well-founded principle to find the appropriate generalization level.

The approach I used to learn selectional preferences is a variant of the Li & Abe approach which is better suited for the task outlined in the introduction. In this section, I describe this method and the experiment I carried out to evaluate its suitability for our task.

2.1 Information theoretic foundations

Information theory deals with coding information as efficiently as possible. In the framework of this discipline, information is usually coded in bits. If one has to code a sequence of signs (in our case, nouns which occur as the complement of a certain verb in a corpus), the simplest way to do this would be to represent each sign by a bit sequence of uniform length. However, if the probabilities of the individual signs differ significantly, it is more efficient (with respect to data compression) to assign shorter bit sequences to more probable (and thus more frequent) signs and longer bit sequences to less probable (and less frequent) signs. It can be shown that one can achieve the shortest average code length by assigning $\lceil log_2 \frac{1}{p(x)} \rceil$ bits to a sign x with probability $p(x)$ (cf. [Cover and Thomas, 1991]). Thus, if one has a good estimation of the probability distribution which underlies the occurrence of the signs, one can develop an efficient coding scheme (a mapping between signs and bit sequences) based on this estimation.

2.2 The tree cut model

The approach in [Abe and Li, 1996] is based on the *Minimum Description Length Principle (MDL)* invented by J. Rissanen (cf. [Rissanen and Ristad, 1992]). This principle is motivated by information theory and is based on the assumption that learning corresponds to data compression: The better one knows which general principles underlie a given data sample, the better one can make use of them to encode this sample efficiently. If one wants to encode a sample, one has to encode (1) the probability model that determines a coding scheme, and (2) the data themselves (by employing that coding scheme). The MDL principle states that the best probability model is that which achieves the highest data compression, i.e. which minimizes the sum of the lengths of (1) (the *model description length*) and (2) (the *data description length*.) In our case, a sample S_v consists of the noun tokens that appear at a certain syntactic argument slot (e.g. the object) of a certain verb in the examined corpus.

Li & Abe represent the selectional behaviour of a verb (with respect to a certain argument) as *tree cut model*. Such a model provides a horizontal cut through the noun hierarchy tree, so that the concepts that are located along the cut form a partition of the noun senses covered by the hierarchy. Each concept is assigned a preference value. The preference value for a concept in the cut is inherited by its subconcepts. A tree cut model (cut + preference values) determines a probability distribution over the sample (see below), and hence a coding scheme. Figure 1 shows an (artificial) example of a tree cut model.

As preference value, Li & Abe estimate the so-called *association norm*:

$$A(nconcept, v) = \frac{p(nconcept, v)}{p(nconcept)p(v)} = \frac{p(nconcept|v)}{p(nconcept)} \tag{1}$$

This measure quantifies the ratio of the occurrence probability of a noun concept $nconcept$ at a certain argument slot[1] of a verb v and the expected occurrence probability of $nconcept$ at this slot if independence between $nconcept$ and v is assumed. This is equivalent to the ratio of the conditional probability of $nconcept$ given v and the probability of $nconcept$ regardless of a particular verb. An association norm greater than 1 indicates preference, an association norm smaller than 1 dispreference of $nconcept$.

Given the marginal probabilities $p(nsense)$ of noun senses (regardless of a particular verb),[2] a tree

[1] This slot is not explicitly referred to in the formula.

[2] These probabilities are also estimated on the basis of a tree cut model by employing the MDL principle; cf. [Abe and Li, 1996] and [Li and Abe, 1995] for details.

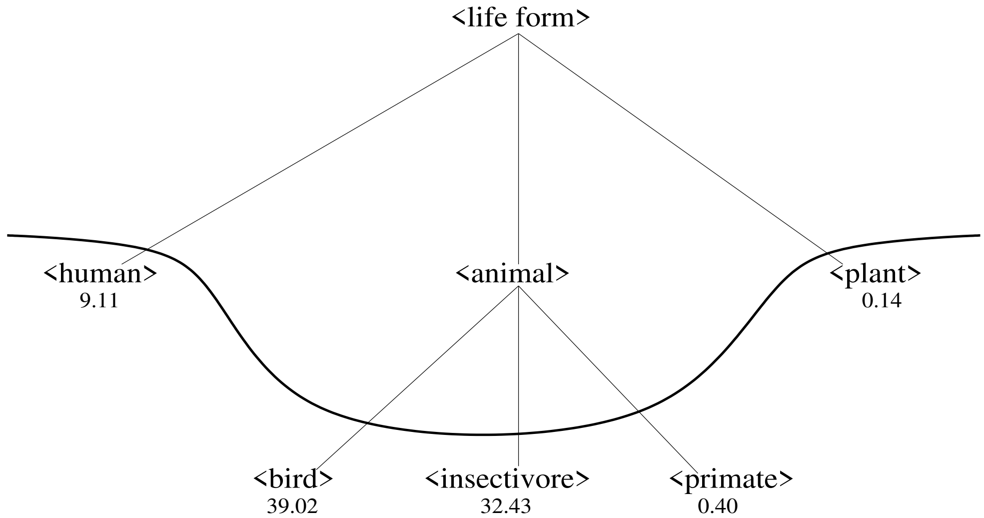

Figure 1: A tree cut model

cut model determines a probability distribution over the noun senses $nsense$ in a sample. This follows from the constraint that the association norm of a concept is inherited by its descendents. Every noun sense $nsense$ is represented in the cut by exactly one concept $nconcept$. So we have

$$p(nsense|v) = A(nsense, v)p(nsense) = A(nconcept, v)p(nsense) \tag{2}$$

The MDL principle is used to get the tree cut model with the appropriate generalization level. The number of bits required to encode a sample S_v using a probability model M is given by

$$L(M) = L_{mod}(M) + L_{dat}(M) \tag{3}$$

with

$$L_{mod}(M) = L_{cut}(M) + L_{par}(M) \tag{4}$$

$L_{mod}(M)$ is the model description length, $L_{dat}(M)$ the data description length, $L_{cut}(M)$ is the code length needed to identify the cut through the hierarchy, and $L_{par}(M)$ is the code length needed to encode the parameters of the model (the association norms of the concepts on the cut). Following the MDL principle, we search for the model M that minimizes $L(M)$.

For simplicity, it is assumed that all possible cuts have uniform probability. Thus, L_{cut} is constant for all cuts. As we aim at minimizing the description length, we can neglect this term.

Li & Abe calculate $L_{par}(M)$ as

$$L_{par}(M) = K\left(\frac{\log|S_v|}{2}\right) \tag{5}$$

K is the number of parameters in M (i.e. the number of concepts on the cut) and $|S_v|$ is the sample size. For every concept on the cut, the association norm is represented by $\frac{\log|S_v|}{2}$ bits. This precision minimizes $L(M)$ for a given M (cf. [Rissanen and Ristad, 1992]).[3]

The data description length is given by

$$L_{dat}(M) = -\sum_{nsense \in S_v} \log p_M(nsense|v) \tag{7}$$

where p_M is the probability distribution determined by M). This equation follows from the information-theoretic principles sketched in section 2.1.

If the tree cut is located near the root, then the model contains only few concepts and the model description length will be low. However, the data description length will be high because the code for the data is based on the probability distribution of the concept in the model, not on the real probability distribution of the noun senses. The coarser the classification is, the more the corresponding distribution p_M deviates from the real distribution. And the more the supposed distribution deviates from the real one, the less efficient the coding scheme, and thus, the longer the code. On the other hand, if the tree cut is located near the leaves, the reverse is true: the fine-grained classification fits the data well, resulting in a low data description length, but the large amount of concepts increases the model description length. Minimizing the sum of these two description lengths yields a balance between compactness (expressing generalization) and accuracy (fitting the data) of the model.

2.3 Introducing a weighting factor

Experiments carried out with the Li & Abe approach revealed a significant drawback of employing the MDL principle for our task (cf. [Wagner, 2000]): The frequency of the examined verb has an undesirable effect on the generalization level of the tree cut model. The algorithm tends to over-generalize (acquire few general concepts) for infrequent verbs and to under-generalize (acquire many specific concepts) for frequent verbs. This behaviour is an immediate consequence of the MDL principle. If a large amount of data has to be described, then the model cost L_{mod} does not contribute much to the whole description length L. The gain of a complex model for encoding the data outweighs the model cost. If, however, only few data have to be captured, then the contribution of L_{mod} to L is much more significant and the cost of encoding a complex model outweighs the gain for encoding the data.

However, this is not the desired behaviour. Generalization should not be triggered by the sample size, but by the "semantic variety" of the instances in the sample: Nouns like "apple", "pear", "strawberry" should generalize to <fruit>. Further instances like "pork" or "cake" should trigger generalization to <food>, and yet further instances like "house" or "vessel" to <physical object>.

The problem discussed above is caused by different complexities of L_{par} and L_{dat} (with respect to the sample size $|S_v|$). As one can see from the equations (5) and (7), L_{par} has the complexity $O(\log|S_v|)$,

[3] For my experiments, I introduced an optimization: It does not make sense to represent the parameter value 0 with $\frac{\log|S_v|}{2}$ bits. (An association norm is 0 if and only if the corresponding concept has no instances in the sample.) A more efficient coding strategy is to mark the concepts that have a non-zero parameter and represent the parameter values for those concepts only. First you need K bits, one for each concept, which indicate whether a concept occurs in the sample or not. Then you need $\frac{\log|S_v|}{2}$ bits for every concept that occurs in the sample. Thus,

$$L_{par}(M) = K + K_{S_v}\left(\frac{\log|S_v|}{2}\right) \tag{6}$$

K_{S_v} is the number of concepts that have instances in the sample. With this modification, one saves $(K - K_{S_v})\frac{\log|S_v|}{2} - K$ bits.

while L_{dat} has the complexity $O(|S_v|)$. Thus, with growing $|S_v|$, L_{dat} "grows faster" then L_{par}, and for frequent verbs, the model description length can be neglected, so that a model with many specific concepts becomes "affordable".

To overcome this drawback, I extended the expression which is to be to minimized by a weighting factor: Instead of minimizing $L_{par} + L_{dat}$, the modified algorithm minimizes

$$L_{par}(M) + C\left(\frac{\log|S_v|}{|S_v|}\right)L_{dat}(M) \quad (C > 0) \tag{8}$$

Now both addends have the same complexity. $|S_v|$ does not directly affect generalization any more.

The value of the constant C influences the degree of generalization. The smaller C is, the more general concepts are acquired. The possibility of manipulating the overall generalization level by the choice of C introduces some flexibility which might prove useful when the algorithm is applied in different situations (tasks, domains, languages, etc.; cf. section 1).

Note that the introduction of weighting is a deviation from the "pure" MDL principle. However, it can be shown that the modified algorithm is consistent with the framework of Bayesian learning (cf. [Quinlan and Rivest, 1989] for a similar approach).

2.4 Evaluation

The experiment described in this subsection was carried out to evaluate the suitability of both the Li & Abe approach (henceforth: standard MDL) and the weighting variant for learning thematic role relations.

2.4.1 The semantic hierarchy

The goal of the learning task is to acquire relations between verb and noun concepts in WordNet. I used WordNet 1.5 for the experiments mentioned in this article. Unlike Li & Abe, I used the algorithms to acquire selectional preferences of verb concepts, not verb forms.

Before being applicable for our learning algorithm, some WordNet specific issues concerning the noun hierarchy had to be addressed. The algorithm requires that the concept hierarchy is a tree where the leaves represent the word senses and the inner nodes represent generalizations of them. However, WordNet is not a pure tree, but a DAG, and all nodes represent both word senses and abstractions of word senses (e.g. the node <person#individual#someone> represents at the same time a general concept denoting all kind of persons *and* a particular word sense for the nouns "person", "individual", and "someone". No hyponym of the node represents this sense. To handle this problem, I introduced for every inner node a virtual leave, i.e. an additional node that captures the noun senses that the node represents, and made this virtual leave a hyponym of the inner node. So all word senses are represented by leaves. To handle the DAG issue, I "broke the DAG into a tree". This means, if a node has more than one parent, I virtually duplicated that node (and its descendants) to maintain a tree structure.[4]

2.4.2 The training data

The training data I used had been compiled at the IMS, University of Stuttgart. These data were retrieved by parsing the British National Corpus with a stochastic parser (cf. [Carroll and Rooth, 1998]).

[4]I will implement a more principled solution to this problem in the future.

Each item in these data consists of a verb, a subcategorization frame, a certain syntactic argument, the head noun of the corresponding complement, and the number of occurrences of this pattern. For example, the item

```
drink#subj:obj:adv/obj  coffee  5
```

represents the pattern that the verb "drink" subcategorizes for a subject, an object, and an adverb, and the object is "coffee". This combination occurs 5 times in the corpus.

This kind of information cannot employed directly for our learning task. One reason for this is that a certain thematic role can have different syntactic realizations. The mapping from syntactic arguments to the underlying semantic complements is all but trivial. As this paper is intended as a case study, I will not systematically address this problem here, but concentrate on instances of the PATIENT role which are realized as an object. Thus, I extracted all verb–object pairs from the data and used them as basis for extracting PATIENT relations.

Another problem is that we want to acquire relations between verb and noun concepts, whereas the training data consist of pairs of verb and noun forms. Hence, these word forms have to be mapped to word senses which correspond to WordNet concepts. To put it in a more formal way, given a verb–noun pair (v, n), you have to estimate the probability $p(sense_i(v), sense_j(n)|v, n)$ for each possible combination $(sense_i(v), sense_j(n))$ of senses of v and n. Lacking further information concerning sense distributions, I assumed

(a) a uniform distribution of the senses of v and n, respectively

(b) independence of the sense distributions of v and n

With these assumptions, the probability of senses given the word forms can be estimated as in equation (9):

$$p(sense_i(v), sense_j(n)|v, n) = \frac{1}{senses(v)} \times \frac{1}{senses(n)} \tag{9}$$

where $senses(w)$ denotes the number of senses of word w. With this information we can estimate the counts of verb–noun sense pairs $\#(sense_i(v), sense_j(n))$ from the counts of verb–noun form pairs $\#(v, n)$ extracted from our training data:

$$\#(sense_i(v), sense_j(n)) = \#(v, n) \times p(sense_i(v), sense_j(n)|v, n) \tag{10}$$

2.5 A gold standard

The learning algorithm is evaluated by comparing the results with a gold standard. There are no thematic relations encoded in WordNet. However, some of the wordnets in EuroWordNet (Dutch, English, Estonian, Italian, and Spanish) contain thematic relations. These relations have been manually encoded or extracted from other lexical resources, respectively. I employed them for compiling a gold standard by mapping them to WordNet.

As noted above, I started from the simplifying heuristic that the patient of a verb is usually syntactically realized as its object. In EuroWordNet, a verb sense is connected to a noun sense that it prefers as

number (percentage) of gold standard concepts	standard MDL		weighting; $C =$ 1,000		10,000	
exactly matched	17	(5.0%)	123	(35.9%)	162	(47.2%)
matched by 1 level hyponym	3	(0.9%)	42	(12.2%)	59	(17.2%)
matched by 1 level hyperonym	13	(3.8%)	39	(11.4%)	44	(12.8%)
matched by $>=$ 2 level hyponym	1	(0.3%)	8	(2.3%)	16	(4.7%)
matched by $>=$ 2 level hyperonym	156	(45.5%)	81	(23.6%)	30	(8.7%)
not matched	153	(44.6%)	50	(14.6%)	32	(9.3%)

Table 1: Comparison of acquired tree cut models with the gold standard

its patient by the INVOLVED_PATIENT relation. Thus, I mapped the relations of this type to WordNet in the following way: I extracted those INVOLVED_PATIENT relations where both the source node and the target node were linked to a node in the inter-lingual index (ILI) by a synonymy or near-synonymy relation. The inter-lingual index essentially consists of all the concept nodes of WordNet 1.5. Thus, extracting the ILI concepts which correspond to the source and the target concept of an INVOLVED_PATIENT relation, respectively, immediately yields a mapping of this relation to WordNet 1.5.

A certain amount of these relations were inappropriate for our task. In some cases, the patient related to a verb is not realized as a direct object (e.g. <silt> INVOLVED_PATIENT <sediment>). In other cases, the patient is represented not extensionally but as an intensional concept (e.g. <address> INVOLVED_PATIENT <addressee>). Such concepts cannot be derived by generalizing from (extensional) corpus instances. To obtain a gold standard that is appropriate for the evaluation of the two algorithms, I excluded these problematic cases. The remaining set contained 662 relations for 368 verb concepts altogether.

2.5.1 Results

As evaluation set, I selected those verb concepts that occurred more than 50 times in the training data. 174 verbal concepts met this criterion.

For both variants standard MDL and weighting (with different values of C), I compared the noun concepts acquired for a verb concept with the corresponding noun concepts in the gold standard. Table 1 shows the number and the percentage of the noun concepts in the gold standard which were exactly matched, not matched at all, or matched by more general or more specific classes in the tree cut model. Dispreferred concepts, i.e. concepts with a preference value < 1, are considered as not matching the gold standard.

While the results for the standard MDL algorithm are all but satisfying, the results for the weighting algorithm are very promising. For $C = 10000$, 47.2% of the noun concepts are exactly matched by the tree cut model (as opposed to 5% with standard MDL). If one also takes the approximate matches (1 level deviation) into account, then the matching rate is 77.2% (vs. 9.7% for standard MDL).[5]

As can be seen, the standard MDL approach tends to learn too general concepts or not to match

[5] It is appropriate to count approximate hits, because human intuition about the *exact* generalization level is not always clear cut as well. For example, the gold standard contains two INVOLVED_PATIENT relation for the verb <own#have#possess>: one to <asset> (originating from the Dutch wordnet) and one to <possession> (originating from the English wordnet). The former concept is a hyponym of the latter one.

verb	weighting model	standard MDL model
<pump#raise_with_a_pump>	<gas>	<entity>
<start#start_up#set_in_motion>	<engine>	—
<climb#climb_up#mount#go_up>	<road#route>	<artifact#artefact>
<send#direct>	<mail>	<relation>
<tame#chasten#subdue>	<animal#beast#creature#fauna>	<entity>
<buy#purchase#take>	<commodity#goods>	<commodity#goods>
<cook#change_by_heating>	<food#nutrient>	<food#nutrient>
<suppress#inhibit#subdue#curb>	<idea#thought>	<psychological_feature>
<record#tape>	<material>	<abstraction>
<operate#control>	<device>	—
<pick#pluck#cull>	<flower#bloom#blossom>	<entity>
<plug_in#connect	<device>	—
<cultivate#foster_the_growth_of>	<plant#flora#plant_life>	—

Table 2: Some gold standard concepts exactly matched by weighting tree cut models ($C = 1000$) and the corresponding concepts in the respective standard MDL tree cut models

preferred concepts at all, respectively. This tendency is illustrated by table 2. This table shows some examples of gold standard concepts which are exactly matched by the respective tree cut models acquired by the weighting algorithm ($C = 1000$), and the corresponding concepts in the respective tree cut models acquired by the standard MDL algorithm. (A dash indicates that a gold standard concept is not matched by the model.)

The best performance is achieved with a rather high value of C, and hence, with a rather low generalization level of the acquired preferences. The reason for this is the nonuniform nature of the gold standard. Unfortunately, the gold standard is very inconsistent with respect to the degree of generalization: On one hand, very general concepts, e.g. <own> INVOLVED_PATIENT <possession>, on the other hand, very specific concepts, e.g. <add> INVOLVED_PATIENT <appendix>, have been encoded. It is obvious that specific concepts are captured by a cut at a low generalization level. More surprisingly, it turned out that in many cases, low-generalization cuts also capture rather general concepts, e.g. <person>. This is due to the treatment of inner nodes sketched in section 2.4.1. It is often the case that a specific cut contains virtual leaves, which represent senses of those words which characterize the corresponding inner nodes. For example, the virtual leave that corresponds to the node <person> represents a sense of the word "person". If a cut contains this leave, then "person" co-occurs with the verb under consideration with a significant frequency in the training sample so that the algorithm recognizes the corresponding concept as preferred. In other words: in such cases, general concepts are acquired due to immediate evidence from the corpus. For these reasons, a bias to low-generalization cuts yields the best overall performance.

Note that this evaluation focuses on recall, i.e. the percentage of relations in the gold standard which are captured by the learned model. Unfortunately, the gold standard is far from being "exhaustive"; in general, it does not represent *all* noun concepts which are preferred by a given verb concept. Furthermore, information about which noun concepts are *dispreferred* by a certain verb concept is not available. Thus, it is not possible to carry out a quantitative evaluation of precision, i.e. the percentage of learned thematic relations which "really" hold.

Manual inspection shows that the acquired selectional preferences contain a considerable amount of noise. For example, the gold standard contains a relation between the verb <hound#hunt#trace> and the noun <game> (an animal hunted for food or sport). The tree cut model for <hound#hunt#trace> correctly models preference for this concept, but additionally for three other senses of "game", e.g. a hyponym of <competition#contest>.

As this example illustrates, one major factor that introduces noise is that the input data are not lexically disambiguated. It is obvious that the two assumptions sketched in section 2.4.2, uniform sense distribution and independence between verb and noun sense, do not hold in reality. Thus, to overcome this drawback, I preprocessed the input data in order to (partially) disambiguate them.

3 Disambiguating the training data

To reduce noise in the acquired selectional preferences, I used a more sophisticated approach for the mapping from forms to senses, i.e. to estimate $p(sense_i(v), sense_j(n)|v, n)$ (cf. section 2.4.2). This approach combines a technique for clustering verb–noun pairs (*latent semantic clustering*, cf. [Rooth et al., 1998]) with a method for disambiguating the verbs and nouns within each cluster by measuring the distance of their respective senses (cf. [Resnik, 1995a]).

3.1 Modeling latent semantic classes

The probability model induced by Rooth et al. aims at clustering verb–noun pairs according to their selectional pattern. For example, verbs like "face", "resolve", "address", or "fight" select nouns like "problem", "pressure", "damage", or "challenge". Such a pattern corresponds to a cluster, a so-called *latent semantic class (LSC)*, that comprises (among others) these verbs and nouns. A latent semantic class is a soft cluster, i.e. class membership is modeled by conditional probabilities. Thus, in general, verbs and nouns are attached to multiple classes, which reflects their polysemy (cf. [Rooth, 1998] for a general motivation of the LSC model).

In the LSC model, the probability of a verb–noun pair (v, n) is expressed as

$$p(v, n) = \sum_c p(c) \times p(v|c) \times p(n|c) \tag{11}$$

where c is a latent semantic class. The parameters of the model, i.e. the probabilities $p(c)$, $p(v|c)$, and $p(n|c)$ for each c, v, and n, are learned by maximum likelihood estimation from incomplete data via the EM algorithm (cf. [Rooth et al., 1998] for details).

For my investigation, the abovementioned training data consisting of verb–noun pairs[6] were used for training the LSC model. The number of classes as well as the number of learning iterations has to be fixed in advance. Following [Rooth et al., 1998], I used 35 classes and 400 iterations.

3.2 Disambiguating verbs and nouns

As illustrated by [Rooth, 1998] and [Rooth et al., 1998] (and confirmed by a manual inspection of the LSC model that I acquired), similar verbs and nouns, respectively, tend to co-occur within one class (or, to be more exact, tend to have high conditional probabilities for the same class). Thus, it makes

[6] Like in [Rooth et al., 1998], the subcategorization information mentioned in section 2.4.2 was attached to the verbs.

sense to employ Resnik's approach for disambiguating semantically similar words in a cluster (cf. [Resnik, 1995a]).

The basic idea of this method is to compare the senses of the words in a cluster and for each word select those sense(s) which is (are) closest to the senses of the other words.

To formalize the notion of "closeness" of senses, a way to measure similarity between senses has to be provided. For this task, Resnik employs the hyponym/hyperonym hierarchy of WordNet. This is based on the assumption that semantic similarity of senses is reflected by their relative position in this hierarchy. Given such a semantic taxonomy, one obvious way to calculate the distance between two senses would be to count the number of edges on the path between these senses. However, the hyperonym relations in WordNet represent different degrees of abstraction, and hence, different semantic distances between the hyponym and the hyperonym (e.g., consider <rabbit_ears> HAS_HYPERONYM <television_antenna> vs. <white_elephant> HAS_HYPERONYM <possession>).

For this reason, Resnik's definition of semantic similarity is based on a corpus-based measure: *information content* (cf. [Resnik, 1995b]). The information content of a concept is defined as

$$info(concept) = -\log p(concept) \tag{12}$$

where $p(concept)$ is estimated as the relative frequency of *concept* in a corpus. The higher the probability of a concept, the less informative (and the more abstract) it is. Resnik defines the similarity of two concepts in WordNet as the amount of information which both concepts share, i.e. the information content of their most informative subsumer.

For each word w in a cluster c, Resnik estimates the probability distribution $p(sense_i|w, c)$ of senses given w and c. To achieve this, he performs a pairwise comparison of the words in c. For each word pair (w_1, w_2), he calculates the similarities of each sense of w_1 with each sense of w_2. For those senses which are in the sense pair(s) with the highest similarity, a counter is incremented by this similarity value. Finally, a normalization transforms the sense counters into sense probabilities. For those words for which no evidence for certain senses can be found, a uniform distribution of the possible senses is assumed.

I applied this method separately for nouns and verbs in each latent semantic class. However, I introduced two modifications. First, I had to take into account that the classes are soft clusters. Thus, I weighted the "support" that one word sense provides for another one by its class membership probability: If a sense pair $sense_{k1}(w_i), sense_{k2}(w_j))$ has the highest similarity for the words w_i and w_j, then the counter for $sense_{k1}(w_i)$ in class c is incremented by $sim(sense_{k1}(w_i), sense_{k2}(w_j)) \times p(w_j|c)$ and vice versa.

Second, the similarity measure used by Resnik does not take into account the distance of the two compared senses from their most informative subsumer. For this reason, I adopted the distance measure defined in [Jiang and Conrath, 1997]. This measure calculates the distance (i.e. the loss of information content) between the two concepts and their most informative subsumer:

$$\begin{aligned} dist(concept_1, concept_2) &= info(concept_1) + info(concept_2) \\ &\quad -2 \times info(mis(concept_1, concept_2)) \\ &= -\log p(concept_1) - \log p(concept_2) \\ &\quad +2 \times \log p(mis(concept_1, concept_2)) \end{aligned} \tag{13}$$

where $mis(concept_1, concept_2)$ denotes the most informative subsumer of $concept_1$ and $concept_2$.

number (percentage) of gold standard concepts	standard MDL		weighting; $C =$ 1,000		10,000	
exactly matched	25	(10.5%)	75	(31.6%)	86	(36.3%)
matched by 1 level hyponym	5	(2.1%)	33	(13.9%)	45	(19.0%)
matched by 1 level hyperonym	19	(8.0%)	24	(10.1%)	25	(10.5%)
matched by >= 2 level hyponym	1	(0.4%)	9	(3.8%)	13	(5.5%)
matched by >= 2 level hyperonym	111	(46.8%)	34	(14.3%)	20	(8.4%)
not matched	76	(32.1%)	62	(26.2%)	48	(20.3%)

Table 3: Comparison of tree cut models acquired from disambiguated data with the gold standard

To be applicable for our task, this distance measure has to be transformed into a similarity measure. This was done by

$$sim(concept_1, concept_2) = 2^{-dist(concept_1, concept_2)} \tag{14}$$

3.3 Estimating sense counts from word form counts

With the probabilities estimated so far, we can estimate the joint probability of a verb sense and a noun sense:

$$\begin{aligned} p(sense_i(v), sense_j(n)) &= \sum_c p(c) \times p(sense_i(v)|c) \times p(sense_j(n)|c) \\ &= \sum_c p(c) \times p(sense_i(v)|v,c)p(v|c) \times p(sense_j(n)|n,c)p(n|c) \end{aligned} \tag{15}$$

The conditional probability of the co-occurrence of a verb sense and a noun sense given the underlying verb and noun form is

$$p(sense_i(v), sense_j(n)|v,n) = \frac{p(sense_i(v), sense_j(n))}{p(v,n)} \tag{16}$$

which is the ratio of the probabilities defined in (15) and (11), respectively. This probability is used to estimate verb–noun sense pair frequencies according to equation (10).

3.4 Evaluation

Again, I selected the verb concepts with a frequency of at least 50 for the evaluation. This selection yielded a set of 122 verbs. Table 3 shows the evaluation results.

Compared to learning from undisambiguated data, the percentage of exactly or approximately (at level 0 or 1) matched concepts is worse for the weighting algorithm, e.g. 65.8% for $C = 10000$ (as opposed to 77.2%). The reason for this is that disambiguation errors misinform the learning algorithm to a certain degree, as they favour "incorrect" senses, which results in dropping "correct" noun concepts contained in the gold standard.[7] However, the results are still promising. Improved approaches to WSD should decrease this effect.

[7] Analogously, some of the "correct" verb senses were dropped. For this reason, less verb concepts in the gold standard are captured by the data.

As expected, disambiguation reduces noise in the tree cut models. For example, two of the three erroneous senses of “game” are not modeled as preferred concepts of <hound#hunt#trace> any more (cf. section 2.5.1).

As already noted in section 2.5.1, there is no way to measure automatically how much noise and how much “good” data is dropped by the disambiguation step. For $C = 10000$, disambiguation yields 573 preferred noun concepts per verb concept on average, as opposed to 1338 concepts without disambiguation. For $C = 1000$, disambiguation reduces this number from 396 to 254. Thus, the relative overall reduction of preferred concepts (57.2% or 35.9%, respectively) is much higher than the loss of accuracy w.r.t. the gold standard. This indicates that noise reduction outweighs the loss of useful information caused by the disambiguation step. To get a better idea about this loss of “correct” information, a detailed manual comparison of the cuts with and without disambiguation will be necessary. Anyway, in a “semi-automatic setting” in which the algorithm learns candidate concepts which are manually inspected afterwards, reducing the “candidate space” without losing too many valid candidates can be very useful.

Standard MDL performs better with disambiguation, but the results are still unsatisfying. Here, for most of the verbs, the tree cut model contains rather general concepts, for which the danger to be “disambiguated away” is low.

4 Conclusion and future work

In this paper, I presented a method for learning thematic role relations for wordnets by means of statistical corpus analysis. This method is based on an approach proposed by [Abe and Li, 1996]. I modified this approach by introducing a weighting factor. An evaluation on a gold standard, which I extracted from EuroWordNet, showed that this method achieves a learning accuracy of up to 77% (whereas the performance of the original approach is low.) I also carried out a preprocessing step for a partial lexical disambiguation of the input data. This disambiguation step substantially reduces the amount of relations proposed by the learning algorithm, accompanied by a comparably moderate loss in performance.

By adapting the weighting factor, it is possible to influence the level of generalization of the learned concepts. This might be useful to fine-tune learning w.r.t. different applications. For example, it is possible to bias the algorithm to learn either general concepts which subsume the preferred concepts in a compact manner, or specific concepts to represent a typical, but not exhaustive set of preferences. In the latter case, additional criteria for adopting the candidate concepts could be imposed, e.g. a threshold for the preference value.

To evaluate the approach on a broader basis, I plan to extend the gold standard by other thematic relations. Furthermore, I will examine different syntactic realizations for these relations. As noted, the current gold standard largely varies w.r.t. the degree of generalization. After its extension, it should be possible to compile sufficiently large subsets which are consistent in this respect. I will evaluate the algorithm on such optimized subsets.

References

[Abe and Li, 1996] Abe, N. and Li, H. (1996). Learning Word Association Norms Using Tree Cut Pair Models. In *Proc. of 13th Int. Conf. on Machine Learning*.

[Agirre and Martinez, 2002] Agirre, E. and Martinez, D. (2002). Integrating selectional preferences in WordNet. In *Proc. of First International WordNet Conference*, Mysore, India.

[Carroll and Rooth, 1998] Carroll, G. and Rooth, M. (1998). Valence induction with a head-lexicalized pcfg. In *Empirical Methods in NLP workshop*, Granada.

[Cover and Thomas, 1991] Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley & Sons, New York.

[Fellbaum, 1998] Fellbaum, C., editor (1998). *WordNet: An electronical lexical database*. MIT Press, Cambridge, Mass.

[Fillmore and Baker, 2001] Fillmore, C. J. and Baker, C. F. (2001). Frame semantics for text understanding. In *Proc. of WordNet and Other Lexical Resources Workshop, NAACL*, Pittsburgh.

[Fillmore et al., 2001] Fillmore, C. J., Wooters, C., and Baker, C. F. (2001). Building a large lexical databank which provides deep semantics. In *Proc. of Pacific Asian Conference on Language, Information and Computation*, Hong Kong.

[Huyck, 2000] Huyck, C. R. (2000). A practical system for human-like parsing. In *Proc. of ECAI 2000*, pages 436–440, Berlin.

[Jiang and Conrath, 1997] Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of International Conference on Research in Computational Linguistics ROCLING X)*, Taiwan.

[Li and Abe, 1995] Li, H. and Abe, N. (1995). Generalizing Case Frames Using a Thesaurus and the MDL Principle. In *Proc. of Int. Conf. on Recent Advances in NLP*.

[Quinlan and Rivest, 1989] Quinlan, J. R. and Rivest, R. L. (1989). Inferring Decision Trees Using the Minimum Description Length Principle. *Information and Computation*, 80:227–248.

[Resnik, 1995a] Resnik, P. (1995a). Disambiguating noun groupings with respect to wordnet senses. In *Proc. of 3rd Workshop on Very Large Corpora*, MIT.

[Resnik, 1995b] Resnik, P. (1995b). Using information to evaluate semantic similarity in a taxonomy. In *Proc. of 14th International Joint Conference on Artificial Intelligence*.

[Resnik, 1993] Resnik, P. S. (1993). *Selection and Information: A Class-Based Approach to Lexical Relationships*. Dissertation, University of Pennsylvania.

[Ribas, 1994] Ribas, F. (1994). An experiment on learning appropriate selectional restrictions from a parsed corpus. In *Proc. of COLING*, Kyoto.

[Rissanen and Ristad, 1992] Rissanen, J. and Ristad, E. S. (1992). Language acquisition in the MDL framework. In Ristad, E. S., editor, *Language Computations*, volume 17 of *Series in Discrete Mathematics and Theoretical Computer Science*, pages 149–166. DIMACS.

[Rooth, 1998] Rooth, M. (1998). Two-Dimensional Clusters in Grammatical Relations. In Rooth, M. et al., editors, *Inducing Lexicons with the EM Algorithm*, volume 4 (3) of *AIMS*, pages 7–24. Universität Stuttgart.

[Rooth et al., 1998] Rooth, M., Riezler, S., Prescher, D., Carroll, G., and Beil, F. (1998). EM-Based Clustering for NLP Applications. In Rooth, M. et al., editors, *Inducing Lexicons with the EM Algorithm*, volume 4 (3) of *AIMS*, pages 98–124. Universität Stuttgart.

[Vossen, 1999] Vossen, P., editor (1999). *EuroWordNet Final Document*. EuroWordNet (LE2-4003, LE4-8328). Deliverable D032D033/2D014.

[Wagner, 2000] Wagner, A. (2000). Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis. In *Proc. of ECAI-2000 Workshop on Ontology Learning*, pages 37–42, Berlin.